Approximate Kernel Selection with Strong Approximate Consistency

Lizhong Ding[1,*], **Yong Liu**[3], **Shizhong Liao**[4], **Yu Li**[2], **Peng Yang**[2],
Yijie Pan[5], **Chao Huang**[6], **Ling Shao**[1], **Xin Gao**[2,*]
[1] Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
[2] King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[3] Institute of Information Engineering, CAS, China, [4] Tianjin University, China
[5] Ningbo Institute of Computing Technology, CAS, China
[6] Ningbo Institute of Information Technology Application, CAS, China

Abstract

Kernel selection is fundamental to the generalization performance of kernel-based learning algorithms. Approximate kernel selection is an efficient kernel selection approach that exploits the convergence property of the kernel selection criteria and the computational virtue of kernel matrix approximation. The convergence property is measured by the notion of approximate consistency. For the existing Nyström approximations, whose sampling distributions are independent of the specific learning task at hand, it is difficult to establish the strong approximate consistency. They mainly focus on the quality of the low-rank matrix approximation, rather than the performance of the kernel selection criterion used in conjunction with the approximate matrix. In this paper, we propose a novel Nyström approximate kernel selection algorithm by customizing a criterion-driven adaptive sampling distribution for the Nyström approximation, which adaptively reduces the error between the approximate and accurate criteria. We theoretically derive the strong approximate consistency of the proposed Nyström approximate kernel selection algorithm. Finally, we empirically evaluate the approximate consistency of our algorithm as compared to state-of-the-art methods. The strong approximate consistency of our algorithm guarantees a consistently better performance than the algorithms with weaker approximate consistency.

Introduction

Kernel-based learning provides a way to implicitly transform data into a new feature space, which allows the learning of nonlinear functions using linear classifiers or regressors in the kernel-induced feature space. The kernel function determines the reproducing kernel Hilbert space (RKHS), which is the hypothesis space of kernel-based learning and hence has an essential influence on the performance of the resulting hypothesis. Therefore, the selection of the kernel function is a central problem in kernel-based learning (Micchelli and Pontil 2005). The problem of kernel selection is closely linked to the generalization error of kernel-based learning algorithms. The kernel with the smallest generalization error is usually regarded as the optimal kernel (Bartlett, Boucheron, and Lugosi 2002; Liu et al. 2017; Liu and Liao 2015). However, in practice one cannot compute the generalization error because the underlying probability distribution of the data is unknown. It is thus common practice to resort to estimates of the generalization error. One can either use empirical estimates that are based on experiments, or theoretical estimates that are based on the upper bounds of the generalization error. Cross-validation (CV) is a well-known empirical estimate of the generalization error. Leave-one-out (LOO), the extreme form of CV, provides an almost unbiased estimate of the generalization error. However, the naïve kernel selection strategy of CV, exhaustive search in the kernel parameter space, is computationally intensive. To speed up CV-based methods, approximate CV approaches were proposed, such as efficient LOO (Cawley and Talbot 2010) and Bouligand influence function CV (BIFCV) (Liu, Jiang, and Liao 2014; Liu et al. 2018). Employing the theoretical estimate bounds of the generalization error as kernel selection criteria is another alternative to experimental methods. Different bound-based criteria introduce different measures of the capacity of the hypothesis space (Bartlett, Boucheron, and Lugosi 2002; Ding et al. 2018), such as Rademacher complexity (Bartlett and Mendelson 2002), local Rademacher complexity (Cortes, Kloft, and Mohri 2013; Li et al. 2018), maximal discrepancy (Bartlett, Boucheron, and Lugosi 2002), maximum mean discrepancy (MMD) (Sriperumbudur et al. 2009; Gretton et al. 2012; Song et al. 2012), covering number (Ding and Liao 2014b) and effective dimensionality (Zhang 2005). However, the computational complexities of the existing kernel selection criteria are at least quadratic in the number of examples l, i.e., $O(l^2)$. This kind of scalability is prohibitive for large-scale problems.

Approximate kernel selection is an emerging and efficient kernel selection approach, which exploits the convergence property of the kernel selection criteria as well as the computational virtue of kernel matrix approximation (Ding and Liao 2011; 2014a; 2017). The basic principle of approximate kernel selection is that it is sufficient to calculate approximate kernel selection criteria, which can discriminate the (nearly) optimal kernel from other candidates with high efficiency. Two theoretical problems are faced by approximate kernel selection: how kernel matrix approximation impacts the kernel selection criterion and whether this impact can be ignored for large enough examples. In (Ding and Liao

[*] Corresponding authors (email: lizhong.ding@inceptioniai.org, xin.gao@kaust.edu.sa)

2014a), the approximate consistency was first defined to theoretically answer these questions, by studying under what conditions and at what speed the approximate kernel selection criterion is close to the accurate one, if at all. It is worth mentioning that the approximate consistency is defined for approximate kernel selection algorithms and different from the classical concept of "consistency" in the learning theory, which is defined for learning algorithms and measures how the learned hypothesis converges to the optimal one that minimizes the expected error in the hypothesis space. The Nyström approximation (Williams and Seeger 2000; Drineas and Mahoney 2005; Yang et al. 2012; Kumar, Mohri, and Talwalkar 2012; Gittens and Mahoney 2013; Jin et al. 2013; Musco and Musco 2017) is a prevailing low-rank matrix approximation method in the machine learning community. However, even for the Nyström approximation with the best kernel matrix approximation error bound (Gittens and Mahoney 2013), it is difficult to prove the strong approximate consistency of the approximate kernel selection method (Ding and Liao 2014a). It has been proven that the best approximate consistency for Nyström methods is the $\frac{1}{2}$-order approximate consistency (weaker than the strong approximate consistency) (Ding and Liao 2014a), which is derived from the Nyström approximation using leverage score sampling (Gittens and Mahoney 2013). Providing the first Nyström approximate kernel selection approach with strong approximate consistency is the goal of this paper.

Sampling distribution is critical to the performance of the Nyström approximation. However, for the existing Nyström methods, the sampling distributions are independent of the specific learning task at hand and focus on the quality of the low-rank matrix approximation, which ignores the performance of the kernel selection criterion used in conjunction with these approximations. The $\frac{1}{2}$-order approximate consistency is the best approximate consistency for Nyström methods (Ding and Liao 2014a), which is likely caused by the isolation between the sampling distribution and the kernel selection. In this paper, we customize an adaptive sampling distribution for the Nyström approximation and propose a novel Nyström approximate kernel selection algorithm with strong approximate consistency. The main contributions of this paper can be summarized as follows. First, a criterion-driven adaptive sampling distribution that iteratively reduces the error between the approximate and accurate criteria is designed for the Nyström approximation for the first time. Second, based on this newly designed sampling distribution, we propose a novel Nyström approximate kernel selection algorithm. Third, we prove the strong approximate consistency of the proposed Nyström approximate kernel selection algorithm. Finally, we conduct empirical evaluations of the approximate consistency of the proposed algorithm as compared to state-of-the-art approximate algorithms.

Related Work

Kernel-based learning algorithms suffer from high computational and storage complexity due to the use of the kernel matrix. Kernel matrix approximation is adopted to effectively reduce the computational and storage burdens of kernel-based learning. To achieve linear complexity in the number of examples, low-rank approximations from subsets of columns are considered, such as the classical Nyström method with different kinds of sampling strategies (Williams and Seeger 2000; Drineas and Mahoney 2005; Zhang and Kwok 2010; Kumar, Mohri, and Talwalkar 2012; Gittens and Mahoney 2013; Musco and Musco 2017), the modified Nyström method (Wang and Zhang 2013), incomplete Cholesky decomposition (Fine and Scheinberg 2002; Bach and Jordan 2005; Bach 2013), sparse greedy approximations (Smola and Schölkopf 2000), and CUR matrix decomposition (Drineas, Mahoney, and Muthukrishnan 2008).

The Nyström method[1] is an effective low-rank matrix approximation method that has been extensively used in different domains of machine learning, such as Gaussian process (Williams and Seeger 2000), spectral grouping (Fowlkes et al. 2004), and manifold learning (Talwalkar, Kumar, and Rowley 2008; Zhang and Kwok 2010). To reduce the matrix approximation error, different sampling strategies for the Nyström approximation have been considered and theoretically analyzed, including uniform sampling (Kumar, Mohri, and Talwalkar 2009), column norm-based sampling (Drineas and Mahoney 2005), k-means clustering sampling (Zhang and Kwok 2010; Si, Hsieh, and Dhillon 2017), and leverage score-based sampling (Gittens and Mahoney 2013). We collectively refer to the above sampling as *fixed sampling*, which determines the distribution of all columns before the sampling procedure. In addition to the fixed sampling, an *adaptive sampling* technique was proposed in (Deshpande et al. 2006), which selects more informative columns in each iteration and iteratively updates the sampling distribution for selecting the next columns. The adaptive sampling has been extended to the Nyström approximation (Kumar, Mohri, and Talwalkar 2012), the modified Nyström approximation (Wang and Zhang 2013) and the ridge leverage score Nyström approximation (Musco and Musco 2017). However, these existing fixed and adaptive sampling distributions are almost independent of the specific learning task at hand. They focus on the quality of the low-rank approximation, the kernel matrix approximation error, rather than the performance of the kernel selection criterion with these approximations.

As compared to (Ding and Liao 2011; 2014a; 2017), this paper designs the first *criterion-driven* adaptive sampling strategy and provides the first approximate kernel selection algorithm with *strong approximate consistency* for classical Nystrom approximation.

Notations and Preliminaries

We use $\mathcal{X}$ to denote the input space and $\mathcal{Y}$ the output domain. Usually we will have $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ for binary classification and $\mathcal{Y} = \mathbb{R}$ for regression. We assume $|y| \leq M$ for any $y \in \mathcal{Y}$, where M is a constant. The training set is denoted by $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\} \in (\mathcal{X} \times \mathcal{Y})^l$. We

[1] In the rest of this paper, if we mention the Nyström method, we refer to the classical Nyström method, not the modified Nyström method. Although the modified Nyström method has a tighter matrix error bound, its computational burden is higher than the classical Nyström method.

consider the Mercer kernel κ in this paper, which is a continuous, symmetric and positive definite function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. The kernel matrix $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{l}$, defined on a finite set of inputs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\} \subseteq \mathcal{X}$ is symmetric and positive definite (SPD). The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$ associated with the kernel κ can be defined as $\mathcal{H}_\kappa = \overline{\text{span}}\{\kappa(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$, and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$ on $\mathcal{H}_\kappa$ is determined by $\langle \kappa(\boldsymbol{x}, \cdot), \kappa(\boldsymbol{x}', \cdot)\rangle_{\mathcal{H}_\kappa} = \kappa(\boldsymbol{x}, \boldsymbol{x}')$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. We use $\|\mathbf{K}\|_2$ and $\|\mathbf{K}\|_{\mathrm{F}}$ to denote the spectral and Frobenius norm of $\mathbf{K}$, respectively. We use $\lambda_t(\mathbf{K})$ for $t = 1, \ldots, l$ to denote the eigenvalues of $\mathbf{K}$ in descending order.

Approximate Kernel Selection

In this section, we first introduce a kernel selection criterion that is very general in supervised learning, and then give a brief review of approximate kernel selection and approximate consistency.

We consider the regularized square loss function, which is very common in the machine learning community,

$$\mathcal{E}(f) = \frac{1}{l}\sum_{i=1}^{l}(f(\boldsymbol{x}_i) - y_i)^2 + \mu\|f\|_{\mathcal{H}_\kappa}^2,$$

where μ denotes the regularization parameter. The optimal function $f_\kappa = \arg\min_{f \in \mathcal{H}_\kappa} \mathcal{E}(f)$. By the representer theorem, we have $f_\kappa = \sum_{i=1}^{l} \alpha_i \kappa(\boldsymbol{x}_i, \cdot)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^{\mathrm{T}} = (\mathbf{K} + \mu l \mathbf{I})^{-1}\boldsymbol{y}$, where $\boldsymbol{y} = (y_1 \ldots, y_l)^{\mathrm{T}}$ and $\mathbf{I}$ denotes the identity matrix. Writing $\mathbf{K}_\mu = \mathbf{K} + \mu l \mathbf{I}$, $\|f_\kappa\|_{\mathcal{H}_\kappa}^2 = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} = \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$. Denoting $\boldsymbol{f}_\kappa = (f_\kappa(\boldsymbol{x}_1), \ldots, f_\kappa(\boldsymbol{x}_l))^{\mathrm{T}}$, we have $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$, which implies

$$\boldsymbol{f}_\kappa - \boldsymbol{y} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y} - \mathbf{K}_\mu\mathbf{K}_\mu^{-1}\boldsymbol{y} = -\mu l \mathbf{K}_\mu^{-1}\boldsymbol{y}.$$

Now, we have

$$\begin{aligned}\mathcal{E}(f_\kappa) &= \frac{1}{l}(\boldsymbol{f}_\kappa - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f}_\kappa - \boldsymbol{y}) + \mu\|f_\kappa\|_{\mathcal{H}_\kappa}^2 \\ &= \mu\boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y}.\end{aligned}$$

$\mathcal{E}(f_\kappa)$ is the regularized empirical error of the optimal function f_κ. For a fixed regularization parameter μ, $\mathcal{E}(f_\kappa)$ only depends on the kernel κ. It is known that κ has a one-to-one correspondence with RKHSs $\mathcal{H}_\kappa$. Different kernels correspond to different RKHSs. In different RKHSs, different optimal functions are derived. We can select the optimal function f_κ that makes $\mathcal{E}(f_\kappa)$ the smallest among all optimal functions, and then the corresponding kernel κ will be the optimal kernel. We define a kernel selection criterion as

$$\mathcal{C}(\mathbf{K}) = \mathcal{E}(f_\kappa) = \mu\boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y}. \tag{1}$$

In the following, we adopt $\mathcal{C}(\mathbf{K})$ as a case to show our Nyström approximate kernel selection algorithm. It is worth noting that our algorithm can be generalized to any other kernel selection criterion $\mathcal{C}()$ that is defined as a function of a kernel matrix. For a prescribed set of kernels $\mathcal{K}$, we can find the optimal kernel by

$$\kappa^* = \operatorname*{arg\,min}_{\kappa \in \mathcal{K}} \mathcal{C}(\mathbf{K}) = \operatorname*{arg\,min}_{\kappa \in \mathcal{K}} \mu\boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y}.$$

There are three cases for the choice of the kernel set $\mathcal{K}$: (i) $\mathcal{K}$ includes a given type of kernel that has finite candidate parameters; (ii) $\mathcal{K}$ includes a given type of kernel that has continuous parameters; (iii) $\mathcal{K}$ is defined as a set of nonnegative combinations of base kernels (Ding and Liao 2017). $\mathcal{C}(\mathbf{K})$ can be applied to these three cases. In this paper, we concentrate on the design of the sampling distribution and the approximate kernel selection algorithm, so we only consider the first case and leave the latter two as future work.

The approximate kernel selection was first studied in (Ding and Liao 2011). Suppose that a kernel selection criterion $\mathcal{C}()$ and a kernel matrix approximation algorithm $\mathcal{A}()$, which uses the training data $\mathcal{S}$ and the kernel κ to generate the approximate matrix $\tilde{\mathbf{K}}$, are given, the approximate kernel selection is developed to select the kernel κ^* as

$$\kappa^* = \operatorname*{arg\,min}_{\kappa \in \mathcal{K}} \mathcal{C}(\mathcal{A}(\mathcal{S}, \kappa)) = \operatorname*{arg\,min}_{\kappa \in \mathcal{K}} \mathcal{C}(\tilde{\mathbf{K}}). \tag{2}$$

Here we denote an approximate kernel selection method $\mathcal{M}$ as a 2-tuple: $\mathcal{M} = (\mathcal{C}(), \mathcal{A}())$. The computational cost for $\mathcal{C}(\mathbf{K})$ defined in (1) is $O(l^3)$, which is prohibitive for big data. The computation of $\mathcal{C}(\tilde{\mathbf{K}})$ could be much more efficient than that of $\mathcal{C}(\mathbf{K})$ due to the specific structure of $\tilde{\mathbf{K}}$. For the Nyström approximation, the Woodbury formula could be used to calculate $\mathcal{C}(\tilde{\mathbf{K}})$ (Ding and Liao 2012) and for the multilevel circulant matrix approximation (Ding and Liao 2017), fast Fourier transform (FFT) could be used.

To demonstrate the rationality of approximate kernel selection, the notion of approximate consistency was defined in (Ding and Liao 2014a), which answers the theoretical questions under what conditions and at what speed the approximate kernel selection criterion converges to the accurate one, if at all[2].

Definition 1. *Suppose we are given an approximate kernel selection method $\mathcal{M} = (\mathcal{C}(), \mathcal{A}())$, where $\mathcal{C}()$ is a kernel selection criterion, and $\mathcal{A}()$ is a kernel matrix approximation algorithm, which uses $\mathcal{S}$ and κ to generate the approximate matrix $\tilde{\mathbf{K}}$. We say the approximate kernel selection method $\mathcal{M}$ is of strong approximate consistency, if*

$$|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \le \varepsilon(l), \tag{3}$$

where $\lim_{l\to\infty} \varepsilon(l) \to 0$. We say $\mathcal{M}$ is of p-order approximate consistency if

$$|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \le \varepsilon(l), \tag{4}$$

where $\lim_{l\to\infty} \varepsilon(l)/l^p \to 0$. There are two scenarios: if $\mathcal{A}$ is a deterministic algorithm, the approximate consistency is defined deterministically; if $\mathcal{A}$ is a stochastic algorithm, (3) *or* (4) *is established under expectation or with high probability.*

Nyström Approximate Kernel Selection

In this section, we materialize the kernel matrix approximation algorithm $\mathcal{A}()$ by the Nyström approximation, cus-

[2] In (Ding and Liao 2014a), the approximate consistency is defined for $\mathcal{A}()$. Here we refine that definition and take approximate consistency as a basic property of $\mathcal{M} = (\mathcal{C}(), \mathcal{A}())$.

tomize an adaptive sampling strategy for the Nyström approximation and propose a novel Nyström approximate kernel selection algorithm with strong approximate consistency.

Suppose we randomly sample c columns of $\mathbf{K}$. Let $\mathbf{C}$ denote the $l \times c$ matrix formed by these columns. Let $\mathbf{D}$ be the $c \times c$ matrix consisting of the intersection of these c columns with the corresponding c rows of $\mathbf{K}$. The Nyström approximation matrix is $\tilde{\mathbf{K}} = \mathbf{C}\mathbf{D}_k^{\dagger}\mathbf{C}^{\mathrm{T}} \approx \mathbf{K}$, where $\mathbf{D}_k$ is the optimal rank k approximation to $\mathbf{D}$ and $\mathbf{D}_k^{\dagger}$ is the Moore-Penrose generalized inverse of $\mathbf{D}_k$.

If we denote the SVD of $\mathbf{D}$ as $\mathbf{D} = \mathbf{U}_{\mathbf{D}}\mathbf{\Sigma}_{\mathbf{D}}\mathbf{U}_{\mathbf{D}}^{\mathrm{T}}$,

$$\mathbf{D}_k^{\dagger} = \mathbf{U}_{\mathbf{D},k}\mathbf{\Sigma}_{\mathbf{D},k}^{\dagger}\mathbf{U}_{\mathbf{D},k}^{\mathrm{T}},$$

where $\mathbf{\Sigma}_{\mathbf{D},k}$ and $\mathbf{U}_{\mathbf{D},k}$ correspond to the top k singular values and singular vectors of $\mathbf{D}$, respectively, then we have

$$\tilde{\mathbf{K}} = \underbrace{\mathbf{C}\mathbf{U}_{\mathbf{D},k}\sqrt{\mathbf{\Sigma}_{\mathbf{D},k}^{\dagger}}}_{\mathbf{V}}\underbrace{\left(\mathbf{C}\mathbf{U}_{\mathbf{D},k}\sqrt{\mathbf{\Sigma}_{\mathbf{D},k}^{\dagger}}\right)^{\mathrm{T}}}_{\mathbf{V}^{\mathrm{T}}},$$

where we let $\mathbf{V} = \mathbf{C}\mathbf{U}_{\mathbf{D},k}\sqrt{\mathbf{\Sigma}_{\mathbf{D},k}^{\dagger}} \in \mathbb{R}^{l \times k}$.

As shown in (2), approximate kernel selection adopts the approximate criterion $\mathcal{C}(\tilde{\mathbf{K}})$ to select the optimal kernel. To calculate the value of $\mathcal{C}(\tilde{\mathbf{K}})$, we need to solve the inverse of $\tilde{\mathbf{K}} + \mu l \mathbf{I}_l$, where $\mathbf{I}_l$ denotes the $l \times l$ identity matrix.

Using the Woodbury formula, we obtain

$$\left(\tilde{\mathbf{K}} + \mu l \mathbf{I}_l\right)^{-1} = \frac{1}{\mu l}\left(\mathbf{I}_l - \mathbf{V}\left(\mu l \mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\right).$$

To solve $(\tilde{\mathbf{K}} + \mu l \mathbf{I}_l)^{-1}\boldsymbol{y}$ in $\mathcal{C}(\tilde{\mathbf{K}})$, we introduce the vector $\boldsymbol{u}$ and let $(\tilde{\mathbf{K}} + \mu l \mathbf{I}_l)\boldsymbol{u} = \boldsymbol{y}$. Then we have

$$\boldsymbol{u} = \frac{1}{\mu l}\left(\boldsymbol{y} - \mathbf{V}\left(\mu l \mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{y}\right). \tag{5}$$

To efficiently solve (5), we introduce a temporary variable

$$\boldsymbol{\omega} : \left(\mu l \mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\boldsymbol{\omega} = \mathbf{V}^{\mathrm{T}}\boldsymbol{y},$$

and then $\boldsymbol{u} = \frac{1}{\mu l}(\boldsymbol{y} - \mathbf{V}\boldsymbol{\omega})$.

We summarize the above computation of $\mathcal{C}(\tilde{\mathbf{K}})$ from Step 9 to Step 14 in Algorithm 1. Before conducting the above approximation procedure, the most important step for the Nyström method is to determine the distribution for sampling c columns from $\mathbf{K}$. The sampling distributions of existing Nyström methods are independent of the specific learning task at hand and mainly focus on the kernel matrix approximation error. The independence between the sampling distribution and the learning task is the main source of the weaker approximate consistency. Here we will customize the sampling distribution of the Nyström approximation for approximate kernel selection. We adopt the adaptive sampling for the Nyström approximation, instead of sampling all columns at one time, to select more informative columns. The adaptive sampling procedure is given by Steps 3 to 8 in Algorithm 1. In each iteration, we only select $s < c$ columns from $\mathbf{K}$[3]. Then according to the errors incurred by these columns we update the distribution that will be used for selecting the next s columns. The error used in the adaptive sampling of (Kumar, Mohri, and Talwalkar 2009) is only the matrix approximation error, which is independent of the learning and kernel selection.

Algorithm 1 Nyström Approximate Kernel Selection

```
Require: Training data $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, candidate kernel set
    $\mathcal{K} = \{\kappa^{(i)} : i \in [N]\}$, number of columns to be chosen $c$,
    initial distribution $\mathcal{D}_0$, number of columns selected at each it-
    eration $s$, regularization parameter $\mu$
Ensure: $\kappa^*$
 1: Initialize: $\mathcal{C}_{\mathrm{opt}} = \infty$;
 2: for each $\kappa \in \mathcal{K}$ do
 3:     Sample $s$ indices according to $\mathcal{D}_0$ to form $\mathcal{I}$;
 4:     $t = c/s - 1$\{Number of iterations\};
 5:     for $i = 1 : t$ do
 6:         $\mathcal{D}_i = \mathrm{UpdateProbability}(\mathcal{I})$;\{Algorithm 2\}
 7:         $\mathcal{I}_i$ = set of $s$ indices sampled according to $\mathcal{D}_i$;
 8:         $\mathcal{I} = \mathcal{I} \cup \mathcal{I}_i$;
 9:     end for
10:     Form $\mathbf{C}$ and $\mathbf{D}$ according to $\mathcal{I}$;
11:     Calculate the SVD of $\mathbf{D}$ as $\mathbf{D} = \mathbf{U}_{\mathbf{D}}\mathbf{\Sigma}_{\mathbf{D}}\mathbf{U}_{\mathbf{D}}^{\mathrm{T}}$;
12:     Let $\mathbf{V} = \mathbf{C}\mathbf{U}_{\mathbf{D},k}\sqrt{\mathbf{\Sigma}_{\mathbf{D},k}^{\dagger}}$;
13:     Solve $\left(\mu l \mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\boldsymbol{\omega} = \mathbf{V}^{\mathrm{T}}\boldsymbol{y}$ to obtain $\boldsymbol{\omega}$;
14:     $\boldsymbol{u} = \frac{1}{\mu l}(\boldsymbol{y} - \mathbf{V}\boldsymbol{\omega})$;
15:     $\mathcal{C}(\tilde{\mathbf{K}}) = \mu \boldsymbol{y}^{\mathrm{T}}\boldsymbol{u}$;
16:     if $\mathcal{C}(\tilde{\mathbf{K}}) \le \mathcal{C}_{\mathrm{opt}}$ then
17:         $\mathcal{C}_{\mathrm{opt}} = \mathcal{C}(\tilde{\mathbf{K}})$;
18:         $\kappa^* = \kappa$;
19:     end if
20: end for
21: return $\kappa^*$;
```

For the kernel selection criterion $\mathcal{C}(\mathbf{K})$, to study the approximate consistency, we need to bound the difference

$$\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}}) = \mu\boldsymbol{y}^{\mathrm{T}}\mathbf{K}_{\mu}^{-1}\boldsymbol{y} - \mu\boldsymbol{y}^{\mathrm{T}}\tilde{\mathbf{K}}_{\mu}^{-1}\boldsymbol{y},$$

where $\tilde{\mathbf{K}}_{\mu} = \tilde{\mathbf{K}} + \mu l \mathbf{I}$. The tighter the difference between $\mathcal{C}(\mathbf{K})$ and $\mathcal{C}(\tilde{\mathbf{K}})$ can be bounded, the stronger the approximate consistency can be established.

In order to reduce the difference between $\mathcal{C}(\mathbf{K})$ and $\mathcal{C}(\tilde{\mathbf{K}})$, we define the error matrix

$$\mathbf{E} = \mathbf{K}_{\mu}^{-1} \otimes \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} - \tilde{\mathbf{K}}_{\mu}^{-1} \otimes \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}},$$

where $\otimes$ is the Hadamard product. This error matrix contains the information of the vector $\boldsymbol{y}$ and the regularized kernel matrix. For each iteration of adaptive sampling, we choose the columns to make the error between the approximate and accurate criteria small. However, calculating the error matrix $\mathbf{E}$ requires computing the inverse of $\mathbf{K}_{\mu}$, which is of $O(l^3)$

[3] We do not calculate the whole kernel matrix $\mathbf{K}$ and then sample. We just calculate the corresponding s columns of $\mathbf{K}$.

Algorithm 2 Update Probability

```
Require: The index set $\mathcal{I}$
Ensure: The distribution $\mathcal{D} = \{p_1, \ldots, p_l\}$
 1: Form $\mathbf{C}$ and $\mathbf{D}$ according to $\mathcal{I}$;
 2: Construct the Nyström approximation matrix $\tilde{\mathbf{K}}$;
 3: $\tilde{\mathbf{E}} = \mathbf{C} \otimes \boldsymbol{y}\boldsymbol{y}_\nu^{\mathrm{T}} - \tilde{\mathbf{C}} \otimes \boldsymbol{y}\boldsymbol{y}_\nu^{\mathrm{T}}$;
 4: for $i = 1 : l$ do
 5:     if $i \in \mathcal{I}$ then
 6:         $p_i = 0$;
 7:     else
 8:         $p_i = \|\tilde{\mathbf{E}}_i\|_2^2$;
 9:     end if
10: end for
11: $p_i = p_i / \sum_{i=1}^{l} p_i$ for $i = 1, \ldots, l$;
12: return $\mathcal{D} = \{p_1, \ldots, p_l\}$;
```

time complexity. We can prove that

$$
\begin{aligned}
&|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \\
=\ & |\mu\boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y} - \mu\boldsymbol{y}^{\mathrm{T}}\tilde{\mathbf{K}}_\mu^{-1}\boldsymbol{y}| \\
=\ & \left|\mu\boldsymbol{y}^{\mathrm{T}}\left(\mathbf{K}_\mu^{-1} - \tilde{\mathbf{K}}_\mu^{-1}\right)\boldsymbol{y}\right| \\
=\ & \left|-\mu\boldsymbol{y}^{\mathrm{T}}[(\mathbf{K} + \mu l\mathbf{I})^{-1}(\mathbf{K} - \tilde{\mathbf{K}})(\tilde{\mathbf{K}} + \mu l\mathbf{I})^{-1}]\boldsymbol{y}\right| \\
\leq\ & \mu\|\boldsymbol{y}^{\mathrm{T}}\|_2\|(\mathbf{K} + \mu l\mathbf{I})^{-1}\|_2\|\mathbf{K} - \tilde{\mathbf{K}}\|_2 \\
& \qquad \|(\tilde{\mathbf{K}} + \mu l\mathbf{I})^{-1}\|_2\|\boldsymbol{y}\|_2 \\
\leq\ & \frac{\mu\|\boldsymbol{y}^{\mathrm{T}}\|_2\|\mathbf{K} - \tilde{\mathbf{K}}\|_2\|\boldsymbol{y}\|_2}{\lambda_{\min}(\mathbf{K} + \mu l\mathbf{I})\lambda_{\min}(\tilde{\mathbf{K}} + \mu l\mathbf{I})} \\
\leq\ & \frac{1}{\mu l^2}\|\boldsymbol{y}^{\mathrm{T}}\|_2\|\mathbf{K} - \tilde{\mathbf{K}}\|_2\|\boldsymbol{y}\|_2.
\end{aligned}
\tag{6}
$$

This upper bound shows that for $\mathcal{C}(\mathbf{K})$ we can reduce the error $\mathbf{K}_\mu^{-1} - \tilde{\mathbf{K}}_\mu^{-1}$ by reducing $\mathbf{K} - \tilde{\mathbf{K}}$. Now we redefine the error matrix as $\mathbf{E} = \mathbf{K}_\mu \otimes \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} - \tilde{\mathbf{K}}_\mu \otimes \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$. Computing $\mathbf{E}$ requires a full pass over $\mathbf{K}_\mu$ which is inefficient for large-scale problems. In order to further reduce the computational burden, we approximate $\mathbf{E}$ as follows

$$\tilde{\mathbf{E}} = \mathbf{C} \otimes \boldsymbol{y}\boldsymbol{y}_\nu^{\mathrm{T}} - \tilde{\mathbf{C}} \otimes \boldsymbol{y}\boldsymbol{y}_\nu^{\mathrm{T}},$$

where $\mathbf{C}$ is the previously sampled columns of $\mathbf{K}$, $\tilde{\mathbf{C}}$ is the corresponding columns of $\tilde{\mathbf{K}}$, $\boldsymbol{y}_\nu$ denotes the first ν elements of $\boldsymbol{y}$ and ν is the number of columns in $\mathbf{C}$. For classification, we regularize the labels to keep the class information: we use $l_+(l_-)$ to denote the number of the positive (negative) data points and let $y_i = 1/l_+$, if $y_i = +1$ and $y_i = -1/l_-$ if $y_i = -1$ for $i = 1, \ldots, l$. For regression, we keep the original labels. The error between $\mathbf{C}$ and $\tilde{\mathbf{C}}$ is always less than the error between $\mathbf{K}_\mu$ and $\tilde{\mathbf{K}}_\mu$, i.e., $\|\mathbf{C} - \tilde{\mathbf{C}}\|_{\mathrm{F}} \leq \|\mathbf{K}_\mu - \tilde{\mathbf{K}}_\mu\|_{\mathrm{F}}$. When we theoretically prove the approximate consistency, we can just derive the related upper bound of $\|\mathbf{K}_\mu - \tilde{\mathbf{K}}_\mu\|_{\mathrm{F}}$. Finally, we define the sampling distribution as

$$\mathcal{D} = \{p_i\}_{i=1}^{l}, \quad p_i = \|\tilde{\mathbf{E}}_i\|_2^2 / \|\tilde{\mathbf{E}}\|_{\mathrm{F}}^2, \quad i = 1, \ldots, l,$$

where $\tilde{\mathbf{E}}_i$ is the i-th column of $\tilde{\mathbf{E}}$. The procedure for updating the sampling distribution is shown in Algorithm 2. Step 5 and Step 6 in Algorithm 2 imply that our sampling is without replacement. The complete Nyström approximate kernel selection algorithm is shown in Algorithm 1. The main computational cost of Algorithm 1 is from Step 10 to Step 12. The time complexity of SVD in Step 10 is $O(c^3)$. In Step 11, the matrix multiplication with $O(lck)$ complexity is conducted. In Step 12, the inverse of $\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)$ is solved by computing its Cholesky factorization with complexity $O(k^3)$. Computing the matrix of the linear system takes $O(lk^2)$ multiplications. The total complexity of Step 12 is thus $O(lk^2)$. Therefore, the total time complexity of Algorithm 1 is $O(N(c^3 + lck))$, where N is the number of candidate kernels, which is linear to the number of examples l. Step 3 of Algorithm 2 can be parallelly done in each column, so the time complexity is $O(l)$.

Before giving the main theorem, we first introduce two assumptions (Ding and Liao 2014a; Jin et al. 2013).

Assumption 1. For $\rho \in (0, 1/2)$ and the rank parameter $k \leq c \ll l$, $\lambda_k(\mathbf{K}) = \Omega(l/c^\rho)$ and $\lambda_{k+1}(\mathbf{K}) = O(l/c^{1-\rho})$, where ρ characterizes the eigen-gap.

Assumption 2. We always assume that the rank parameter k is a constant and the sampling size c is a small ratio r of l.

Assumption 1 states the large eigen-gap in the spectrum of $\mathbf{K}$ (Jin et al. 2013), i.e., the first few eigenvalues of $\mathbf{K}$ are much larger than the remaining ones, which is not a strong assumption. As assumed in (Bach 2013), the eigenvalues of the kernel matrix have polynomial or exponential decay. The eigenvalues of Gaussian kernels have exponential decay (Cortes, Kloft, and Mohri 2013). Assumption 1 is always weaker than the exponential decay, even when ρ goes to 0. When ρ is close to 1/2, Assumption 1 is weaker than the polynomial decay. The assumption of the constant rank has been adopted in (Wang and Zhang 2013). Assumption 2 is one of the common settings for the Nyström approximation. The following theorem shows the strong approximate consistency of Algorithm 1, whose proof sketch is given in the Appendix.

Theorem 1. *If Assumptions 1 and 2 hold, for the kernel selection criterion $\mathcal{C}(\mathbf{K})$ defined in* (1)*, we have*

$$\mathbb{E}\left(|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})|\right) \leq \varepsilon(l),$$

where the calculation of $\mathcal{C}(\tilde{\mathbf{K}})$ is shown in Algorithm 1,

$$\varepsilon(l) = \tau\sqrt{\frac{k(k+1)M^8 + sM^4}{s\mu^2 r^{2-2\rho}}}\frac{\sqrt{l-k}}{l^{1-\rho}}$$

for some constant τ and $\lim_{l\to\infty}\varepsilon(l) \to 0$.

The strong approximate consistency reveals the fast convergence of the difference between the approximate kernel selection criterion and the accurate one. When our adaptive Nyström approximate kernel selection algorithm is applied to the kernel selection problem, we can obtain the optimal kernel that is closest to the one produced by accurate kernel selection methods as compared to other Nyström approximate algorithms with weaker approximate consistency,

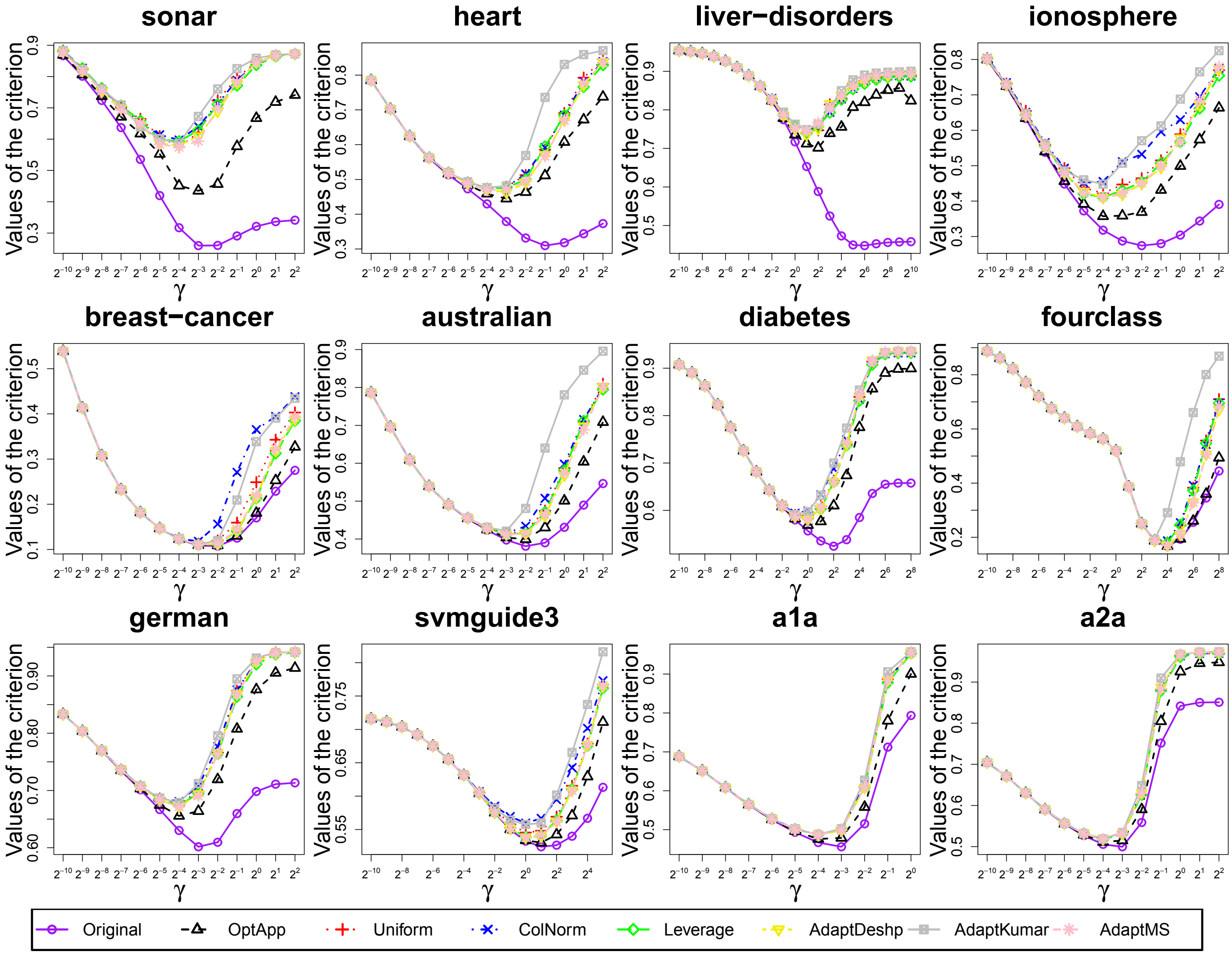

Figure 1: Approximate consistency for different kernel matrix approximation algorithms. These figures show the values of the criteria on different kernel parameters γ.

which shows the appositeness of our adaptive Nyström approximate algorithm for kernel selection.

Empirical Studies

Here we empirically evaluate the approximate consistency of the proposed approximate kernel selection algorithm.

We compare 8 different methods. The first one adopts the original criterion $\mathcal{C}(\mathbf{K})$ in (1) for kernel selection (Original). The second is the optimal rank k approximation (OptApp). For fixed sampling distributions, we consider the Nyström approximation with uniform sampling (Uniform), column norm-based sampling (ColNorm) (Drineas and Mahoney 2005), and leverage score-based sampling (Leverage) (Gittens and Mahoney 2013). For adaptive sampling distributions, we compare the adaptive sampling in (Deshpande et al. 2006) (AdaptDeshp) and the one in (Kumar, Mohri, and Talwalkar 2012) (AdaptKumar). We denote our sampling strategy as "AdaptMS" which is short for model selection-based adaptive sampling. We set the sampling size $c = 0.2l$ and the adaptive sampling size $s = 0.1c$. To avoid the randomness, we run all methods 10 times. Since Gaussian kernels are universal (Steinwart 2002), we adopt Gaussian kernels $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$ with variable width γ as our candidate kernel set $\mathcal{K}$. The focus of this paper is not on tuning the regularization parameter μ, so we just set $\mu = 0.005$. Since the regularized kernel matrix $\mathbf{K}_\mu = \mathbf{K} + \mu l\mathbf{I}$, $\mu = 0.005$ is not too small. All the implementations are in the R language. We conduct experiments on benchmark data sets from UCI repository[4] and LIBSVM Data[5]. Since the aim of our experiments is to evaluate the theoretical findings on approximate consistency, we do not conduct experiments on very large scale datasets.

For each kernel parameter γ, we generate the kernel matrix $\mathbf{K}$ and then use different approximation methods to produce the approximate kernel matrices $\tilde{\mathbf{K}}$. We compare the values of $\mathcal{C}(\mathbf{K})$ and $\mathcal{C}(\tilde{\mathbf{K}})$. The results are shown in Figure 1. We can see that, apart from the optimal rank k approximation (OptApp), the curves of "AdaptMS" are closest to

[4] http://www.ics.uci.edu/∼mlearn/MLRepository.html

[5] http://www.csie.ntu.edu.tw/∼cjlin/libsvm

the curves of the original criterion $\mathcal{C}(\mathbf{K})$ for all data sets, which shows stronger approximate consistency as compared to other approximation methods. It is worth noting that the complexity of OptApp is $O(l^3)$ for each candidate kernel, whereas the complexity of our algorithm is $O(c^3 + lck)$ $(k \leq c \ll l)$. These results demonstrate that when we conduct approximate kernel selection, our Nyström approximate kernel selection algorithm can obtain the kernel that is closest to the one produced by the accurate kernel selection algorithm as compared to the approximate algorithms with weaker approximate consistency.

The complexities of Original and OptApp are all $O(l^3)$. Our adaptive Nyström approximate kernel selection algorithm is faster than OptApp, Leverage and AdaptDeshp, close to AdaptKumar, but slower than Uniform and ColNorm. AdaptDeshp requires a full pass through $\mathbf{K}$ in each iteration. Leverage requires SVD to compute the sampling distribution with $O(l^3)$ time complexity. Although there are cases where some other approximate algorithms have comparable performance to our proposed algorithm, the strong approximate consistency of our algorithm can guarantee a consistently better performance than the algorithms with weaker approximate consistency.

Conclusions

In this paper, we proposed a novel Nyström approximate kernel selection method. By introducing a criterion-driven adaptive sampling distribution, we established the first strong approximate consistency of the Nyström approximate kernel selection method. The sampling strategy considered in this paper is different from the existing matrix-error-based sampling strategies and closely related to the specific learning task at hand. This design for the sampling distribution may open a door for the research into learning-error-based kernel matrix approximation. Through empirical studies, we showed the stronger approximate consistency of the proposed adaptive Nyström approximate kernel selection method as compared to the state-of-the-art algorithms. In future, we consider the application of our theoretical and algorithmic results into online learning (Yang, Zhao, and Gao 2018) and recommendation (Yang et al. 2018).

Appendix: Proof Sketch of Theorem 1

The proof is mainly based on the results in (Deshpande et al. 2006; Cortes, Mohri, and Talwalkar 2010; Deshpande and Rademacher 2010). According to (6), we can bound

$$|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \leq \frac{1}{\mu l^2} \|\boldsymbol{y}^{\mathrm{T}}\|_2 \|\mathbf{K} - \tilde{\mathbf{K}}\|_2 \|\boldsymbol{y}\|_2,$$

Since $\|\boldsymbol{y}\|_2 \leq \sqrt{l}M$, we have

$$|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \leq \frac{M^2}{\mu l} \|\mathbf{K} - \tilde{\mathbf{K}}\|_2.$$

As discussed in (Gittens and Mahoney 2013), SPSD matrix approximations based on column sampling (such as Nyström method) and those based on mixtures of columns can both be subsumed in the *SPSD Sketching Model*. Here we will apply the results for the approximation on mixtures of columns to the Nyström approximation. The initial s columns are sampled according to the efficient volume sampling (Deshpande and Rademacher 2010). We use $\mathbf{C}_s$ to denote the matrix formed by the s columns of $\mathbf{K}$ and $\mathbf{D}_s$ to denote the corresponding intersection matrix. From Theorem 8 in (Deshpande and Rademacher 2010), we obtain

$$\|\mathbf{K} - \mathbf{C}_s \mathbf{D}_{s,k}^{\dagger} \mathbf{C}_s^{\mathrm{T}}\|_{\mathrm{F}} \leq \sqrt{k+1} \|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}}.$$

According to Theorem 2.1 in (Deshpande et al. 2006), for adaptive sampling, if we further adaptively sample another s columns, we have

$$\mathbb{E}\left(\|\mathbf{K} - \mathbf{C}_{2s} \mathbf{D}_{2s,k}^{\dagger} \mathbf{C}_{2s}^{\mathrm{T}}\|_{\mathrm{F}}^2 \right) \leq \|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}}^2 + \frac{k}{s} \|\mathbf{E}\|_{\mathrm{F}}^2.$$

Here $\tilde{\mathbf{K}}$ is $\mathbf{C}_{2s} \mathbf{D}_{2s,k}^{\dagger} \mathbf{C}_{2s}^{\mathrm{T}}$. If we continue the adaptive sampling, the error will decrease. From the theoretical perspective, we just need to bound the sampling twice. Since $\|\boldsymbol{y}\|_2 \leq \sqrt{l}M$, we can prove that

$$\|\mathbf{E}\|_{\mathrm{F}}^2 \leq M^4 \|\mathbf{K} - \mathbf{C}_s \mathbf{D}_{s,k}^{\dagger} \mathbf{C}_s^{\mathrm{T}}\|_{\mathrm{F}}^2.$$

Now, we have

$$\begin{aligned} &\mathbb{E}\left(\|\mathbf{K} - \mathbf{C}_{2s} \mathbf{D}_{2s,k}^{\dagger} \mathbf{C}_{2s}^{\mathrm{T}}\|_{\mathrm{F}}^2 \right) \\ \leq\ &\frac{k(k+1)M^4 + s}{s} \|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}}^2. \end{aligned}$$

Since $\|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}} \leq \sqrt{l-k}\lambda_{k+1}(\mathbf{K})$ and the fact that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathrm{F}}$, according to the above derived bounds, Assumption 1 and Assumption 2, we can obtain $|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\tilde{\mathbf{K}})| \leq \varepsilon(l)$ with

$$\varepsilon(l) = O\left(\sqrt{\frac{k(k+1)M^8 + sM^4}{s\mu^2 r^{2-2\rho}}} \frac{\sqrt{l-k}}{l^{1-\rho}} \right).$$

Since $\rho < \frac{1}{2}$, $\lim_{l \to \infty} \varepsilon(l) \to 0$.

Acknowledgments

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3007-01-01 and BAS/1/1624-01-01, National Natural Science Foundation of China (No. 61673293), National Natural Science Foundation of China (No. 61703396) and Shenzhen Government (GJHZ20180419190732022).

References

Bach, F. R., and Jordan, M. I. 2005. Predictive low-rank decomposition for kernel methods. In *ICML*, 33–40.

Bach, F. 2013. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, 185–209.

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Bartlett, P. L.; Boucheron, S.; and Lugosi, G. 2002. Model selection and error estimation. *Machine Learning* 48(1–3):85–113.

Cawley, G. C., and Talbot, N. L. C. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11:2079–2107.

Cortes, C.; Kloft, M.; and Mohri, M. 2013. Learning kernels using local Rademacher complexity. In *NIPS 26*, 2760–2768.

Cortes, C.; Mohri, M.; and Talwalkar, A. 2010. On the impact of kernel approximation on learning accuracy. In *AISTATS*, 113–120.

Deshpande, A., and Rademacher, L. 2010. Efficient volume sampling for row/column subset selection. In *FOCS*, 329–338.

Deshpande, A.; Rademacher, L.; Vempala, S.; and Wang, G. 2006. Matrix approximation and projective clustering via volume sampling. *Theory of Computing* 2:225–247.

Ding, L., and Liao, S. 2011. Approximate model selection for large scale LSSVM. *Journal of Machine Learning Research - Proceedings Track* 20:165–180.

Ding, L., and Liao, S. 2012. Nyström approximate model selection for LSSVM. In *PAKDD*, 282–293.

Ding, L., and Liao, S. 2014a. Approximate consistency: Towards foundations of approximate kernel selection. In *ECML PKDD*, 354–369.

Ding, L., and Liao, S. 2014b. Model selection with the covering number of the ball of RKHS. In *CIKM*, 1159–1168.

Ding, L., and Liao, S. 2017. An approximate approach to automatic kernel selection. *IEEE Transactions on Cybernetics* 47(3):554–565.

Ding, L.; Liao, S.; Liu, Y.; Yang, P.; and Gao, X. 2018. Randomized kernel selection with spectra of multilevel circulant matrices. In *AAAI*, 2910–2917.

Drineas, P., and Mahoney, M. W. 2005. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research* 6:2153–2175.

Drineas, P.; Mahoney, M. W.; and Muthukrishnan, S. 2008. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications* 30(2):844–881.

Fine, S., and Scheinberg, K. 2002. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research* 2:243–264.

Fowlkes, C.; Belongie, S.; Chung, F.; and Malik, J. 2004. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2):214–225.

Gittens, A., and Mahoney, M. W. 2013. Revisiting the Nyström method for improved large-scale machine learning. In *ICML*, 567–575.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13(1):723–773.

Jin, R.; Yang, T.; Mahdavi, M.; Li, Y.-F.; and Zhou, Z.-H. 2013. Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory* 5(10):6939–6949.

Kumar, S.; Mohri, M.; and Talwalkar, A. 2009. Sampling techniques for the Nyström method. In *AISTATS*, 304–311.

Kumar, S.; Mohri, M.; and Talwalkar, A. 2012. Sampling methods for the Nyström method. *Journal of Machine Learning Research* 13:981–1006.

Li, J.; Liu, Y.; Yin, R.; Zhang, H.; Ding, L.; and Wang, W. 2018. Multi-class learning: from theory to algorithm. In *NIPS 31*.

Liu, Y., and Liao, S. 2015. Eigenvalues ratio for kernel selection of kernel methods. In *AAAI*, 2814–2820.

Liu, Y.; Liao, S.; Lin, H.; Yue, Y.; and Wang, W. 2017. Infinite kernel learning: generalization bounds and algorithms. In *AAAI*, 2280–2286.

Liu, Y.; Lin, H.; Ding, L.; Wang, W.; and Liao, S. 2018. Fast cross-validation. In *IJCAI*, 2497–2503.

Liu, Y.; Jiang, S.; and Liao, S. 2014. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In *ICML*, 324–332.

Micchelli, C. A., and Pontil, M. 2005. Learning the kernel function via regularization. *Journal of Machine Learning Research* 6:1099–1125.

Musco, C., and Musco, C. 2017. Recursive sampling for the Nyström method. In *NIPS 30*, 3836–3848.

Si, S.; Hsieh, C.-J.; and Dhillon, I. S. 2017. Memory efficient kernel approximation. *Journal of Machine Learning Research* 18:1–32.

Smola, A. J., and Schölkopf, B. 2000. Sparse greedy matrix approximation for machine learning. In *ICML*, 911–918.

Song, L.; Smola, A.; Gretton, A.; Bedo, J.; and Borgwardt, K. M. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13:1393–1434.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Lanckriet, G. R. G.; and Schölkopf, B. 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS 22*, 1750–1758.

Steinwart, I. 2002. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* 2:67–93.

Talwalkar, A.; Kumar, S.; and Rowley, H. 2008. Large-scale manifold learning. In *CVPR*, 1–8.

Wang, S., and Zhang, Z. 2013. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research* 14:2729–2769.

Williams, C. K. I., and Seeger, M. 2000. Using the Nyström method to speed up kernel machines. In *NIPS 13*, 682–688.

Yang, T.; Li, Y.-F.; Mahdavi, M.; Jin, R.; and Zhou, Z.-H. 2012. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *NIPS 25*, 1060–1068.

Yang, P.; Zhao, P.; Zheng, V. W.; Ding, L.; and Gao, X. 2018. Robust asymmetric recommendation via min-max optimization. In *SIGIR*, 1077–1080.

Yang, P.; Zhao, P.; and Gao, X. 2018. Bandit online learning on graphs via adaptive optimization. In *IJCAI*.

Zhang, K., and Kwok, J. T. 2010. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks* 21(10):1576–1587.

Zhang, T. 2005. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation* 17(9):2077–2098.